DATA COLLECTION AND ANALYSIS IN THE AIR TRAVEL PLANNING DOMAIN

Jacqueline C. Kowtko, Patti J. Price
Speech Research Program, SRI International, Menlo Park, CA 94025

ABSTRACT

We have collected, transcribed and analyzed over 8 hours of human-human interactive problem solving dialogue in the air travel planning domain, including traveler-agent dialogues and the more constrained agent-airline dialogues. We have used this data to define and test an initial vocabulary, and to design an appropriate interface for the air travel planning domain. The initial interface design was tested via simulation, using 44 subjects solving air travel problems. Our data analysis reveals great differences between the traveler-agent interactions and the agent-airline interactions, with the traveler-simulation interactions falling somewhat in between.

INTRODUCTION

Spoken language systems must, obviously, deal with spontaneous speech. However, most research to date has dealt primarily with read speech, because read speech is much easier to collect in a controlled manner. There are, however, substantial differences between read speech and spontaneous speech. Differences include the many phenomena that are less likely to occur in read speech (pauses, speech and grammatical false starts, filler words, non-standard grammar), as well as important phonological phenomena, such as the frequency of deletions (Bernstein and Baldwin, 1985). On the other hand, it is possible that both the speech and the language of human-machine interactions in a restricted domain will be more constrained and more predictable than those occurring in human-human spontaneous interactions. The goal of the preliminary work presented here is to collect and analyze spontaneous, goal-directed speech and language in the interest of designing and evaluating eventual spoken language systems.

Perhaps the greatest variable affecting performance in current and future systems is the human involved in the human-machine interface. It is therefore important to assess systems over many different subjects. We have chosen the domain of air travel planning because it provides a natural problem-solving domain familiar to many people (120 SRI employees per day on average use spoken interactions to solve travel planning problems). This has greatly facilitated the task of collecting data. Further, the domain can be constrained as desired for initial development (as we have done by allowing only one-way travel between two cities), or expanded naturally to include a great deal of complex problem-solving for future SLSs (inclusion of data on connections, classes of seats, and restrictions on fares, availability of fares, hotels, car rentals, expert system reasoning, etc.). In addition, the air travel planning domain has the advantage of large, real databases in the public domain.

We initially studied human-human interactions, to gain insight into how interactive problem solving is currently used in this domain. We noted that database queries were rare, and that more typically the traveler expresses a few constraints, and then the agent takes the lead and asks questions. We wondered how adaptable subjects would be in a simulated machine interaction: would their travel planning task be more difficult if they were forced to use only database queries? We, simulated an SLS in two conditions: one that permitted the expression of constraints but that were not strictly database queries ("I

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **1989**	2. REPORT TYPE	3. DATES COVERED **00-00-1989 to 00-00-1989**

4. TITLE AND SUBTITLE **Data Collection and Analysis in the Air Travel Planning Domain**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **SRI International,333 Ravenswood Avenue,Menlo Park,CA,94025**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S) 11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES **7**	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**			

need to be there before 3 pm"), and one which accepted only database queries (responding "cannot handle that request" to any other type of utterance). The system responds, in both conditions, with graphics placed on the user's screen (shared information, schedule tables, fare tables, etc.).

The goal of this initial work is to assess human-human problems solving in the air travel domain, and to assess possible differences between human-human and human-machine interactions. It is clear that people are very adaptable, far more so than our current technology. It is not so clear how adaptable they will be and on what dimensions in human-machine interactions. What aspects of the interaction will require a technological solution and what aspects can be handled via a human factors solution? If, for example, it is desirable to handle only database queries, how difficult is it for humans to adapt to this restriction? This is but one example of a myriad of similar questions that could be asked using such simulations. The answers to these questions will expedite the design of efficient human-machine collaborative systems.

METHOD

Before collecting data from human-machine interactions, we observed problem solving in human-human dialogues. Human-human dialogues provide some knowledge of subjects' expectations of the system, the problems which could arise, and solution paths subjects might choose.

Human-human data collection

We collected more than 12 hours (over 100 conversations) of on-site tape recordings of 6 travel agents at a travel agency interacting with clients and with airline agents via telephone. Tape recording equipment was out of the sight of the agent. Both parties knew their voices were being recorded. However, after a few brief interchanges, conversations proceeded as usual. Data collection occurred at the busiest time of day. The tape recorder stayed on for 45-minute durations, except when personal calls interrupted. For each reservation a client makes, agents estimated that the client calls an average of three times: to ask information, to book a flight, and to ticket the flight or make slight changes. We were most interested in first-time calls in which clients booked a flight, although we included data from all three types of calls in our analysis.

Human-machine data collection

To simulate an air travel planning spoken language system, we combined a database retrieval program and a human speech-recognizer/database-accessor, the "wizard." The experiments involved two computer consoles. One Sun 4 graphics console displayed three windows for the subject: a template window of shared information (fields for departure city, arrival city, date, earliest departure time, latest departure time, earliest arrival time, and latest arrival time), a flights schedule window, and a fare window. The wizard could also send a limited number of messages to the subject: "Cannot handle that request", "Would you please repeat that?", and "Ready for more speech input." The subject's console was controlled by the wizard's Sun 3 console, in another office down the hallway. The wizard entered data into the database retrieval program by clicking the mouse.

The user wore a Sennheiser headset microphone, connected to a tape recorder, and spoke to the system via an unobtrusive speakerphone. The system's only means of response was through graphic display. A two-pitch tone coming from the telephone before and after each condition indicated that the experimental system was turned either on or off.

A current total of 44 subjects (26 men, 18 women) participated in the simulated human-machine interactive experiment. Electronic failure caused the loss of data from one (male) subject, leaving 43 who successfully completed their tasks. Two travel planning tasks (one more constrained by fare and the other by schedule, described further below) were assigned each subject counter-balanced with two interaction conditions (database queries only or "regular" -- expressing constraints such as "I can't leave till 3 pm" allowed). The order cycled every four subjects. One quarter of the subjects participated in each of the following test orders:

1. fare task in database query condition, schedule task in regular condition,
2. schedule task in regular condition, fare task in database query condition,
3. fare task in regular condition, schedule task in database query condition, and
4. schedule in database query condition, fare task in regular condition.

Subjects were presented with general written instructions indicating that they were going to help assess and debug an experimental computer-aided travel planner using voice input. Whether the system was completely automated or not was purposefully left ambiguous. The experimenter, the same person as the wizard (author JK), always referred to the experimental system as "the system" or "it." The subject was asked to make a simple flight reservation, interacting with the system to find an optimal flight for the assigned task. General examples of acceptable and unacceptable utterances were provided. The subject was requested to end the session by saying, "ok, book that one." The subject was also told that as the system received information, it would begin to display pieces of information in the template display window. The experimenter then read instructions describing the assigned travel-planning task to the subject, allowing the subject to take notes. This was to avoid any poisoning of the data that might be induced if the subjects simply read the task description. The experimenter then explained the condition to the subject (database query only or regular). Examples of acceptable and unacceptable database queries were given for the relevant condition, and the idea that a database query is a sentence that results in a database retrieval was explained. The subject was also told what types of information the system could provide. The tasks, which each took about 5 minutes to complete, are described below :

A. Book a one-way flight from San Francisco to Los Angeles, for <date>, leaving after <time>, arriving before <time>, subject to the following ordered constraints:
   1. cost under $200
   2. arrive as early as possible (after <time>)
   3. prefer SFO airport to OAK or SJC, and prefer LAX to Burbank

B. Book a one-way flight from San Francisco to Los Angeles, for <date>, arriving before <time>, leaving after <time>, subject to the following ordered constraints:
   1. arrive as close as possible to <time>
   2. spend as little time in transit as possible
   3. prefer SJC airport departure to SFO or OAK
   4. price under $400

The flight information database used is a subset of the Official Airline Guide (OAG) database obtained from the OAG in May 1989. The data was reformatted to allow for easier access and to avoid infringing on OAG's proprietary rights in any later distribution of the data. The data was accessed via a wizard's interface. Developing tools for the wizard is an important task. The wizard takes complete control of the speech and natural language functions of the system and needs a swift means of retrieving data for the user. Being the wizard is difficult because the human must simulate the consistent and more limited response of a computer. By accepting an utterance or producing an error message, the wizard has a large influence over the user's expectation of the system's capabilities.

The wizard accessed the database upon request from the user and controlled the screen of the user by showing tables of fares and schedules, displaying an error message, or requesting that the user ask another question or repeat the previous question. The wizard's screen displayed the same three windows as the subjects' and had additional windows for inputting information with the mouse. The mouse was used to select a category such as departure city and then select the proper value from a pop-up window. The wizard's screen always showed a superset of the information displayed on the user's screen.

RESULTS AND ANALYSIS

The recorded data was first transcribed and verified. Then, various phenomenon that might characterize differences between the styles and conditions examined were counted: number of words, new vocabulary items (items not seen in any previous data), and number of "um"s and other pause fillers. For the human-machine interaction, we also analyzed grammatical false starts ("show me the how many fares are under $200") and speech false starts ("sh- show me only the ones under $200").

Human-Human Data

Twelve hours of data were recorded and transcribed. Of them, 8 hours were verified and analyzed for various characteristics including those in the table below. Note that "naive" user refers to the traveler in the traveler to travel agent conversations and "expert" user refers to the more constrained speech of the travel agent to the airline agent:

User	# Dialogues	# Words	Vocab	# "um"	% um
naive	48	9,315	1,076	501	5.4
expert	10	737	230	21	2.8

Experience is a major factor in dialogue efficiency. Compare the 194 words per dialogue for "naive" users to the 74 words per dialogue for the experts. The vocabulary size also changes significantly between types of user, though this is more difficult to assess given the smaller data set. However, our intuitions, based on looking at these data, is that the vocabulary is substantially more restricted for the agent-agent dialogues for two reasons: the travel agent does not try to gain the sympathy of the airline agent (which travelers often do and which opens up the vocabulary tremendously), and both agents know very well what the other can do (which reduces the vocabulary significantly). Humans interacting with machines will not be likely to try to gain the machine's sympathy, but they will use a much larger vocabulary than otherwise if they are unsure about just what capabilities the system has. We have observed this phenomena in our human-machine simulations. Another measure of efficiency is the frequency of pause fillers , which differs in the two conditions by a factor of 2. Expert users are more concise, following a well-practiced script. Both parties have a clear idea of what each can do for the other and both want an efficient, brief conversation. Pause fillers occur in these conversations primarily when the conversation is focused on new or unknown material such as a client's seat number or an unusual regulation. In the human-human data, when the traveler is unsure of the capabilities of the the agent, the agent takes an active role in guiding the traveler. Interactive conversation, as opposed to one-way communication, increases the efficiency of problem-solving (Oviatt & Cohen, 1988). This will likely be important in designing efficient SLSs for naive, untrained users.

We classified 30 conversations from the data in terms of general type of query used. Five of the 30 conversations were database query-oriented; most of the observed were not strictly database queries, but, rather, expressed constraints related to the problem to be solved . Four of the five database style conversations are from information-only calls, where no booking was made. Information calls from the human-human transcripts usually don't involve all pieces of information necessary for booking a trip. In many cases the traveler merely wants airfare for a trip from X to Y on day Z. Specific flight information and seating arrangements are left for later.

In assessing the design of initial vocabulary, we took 10 dialogues, filled out the items syntactically and semantically, and added a list of function words we had for other purposes. The percent of new words observed in each successive dialogue (where those observed are added to the pool) declines substantially as new dialogues are included. It does not, however, appear to dip below about 3% even after 48 dialogues. This is not a surprising result; it only highlights the need for dealing with (detecting, forming speech models, syntactic models and semantic models for) words outside the expected vocabulary.

Human-Machine Data

We ran two air travel planning sessions per subject. There were two separate tasks as described above, crossed with two query styles: database query and "regular" (expressing constraints). Compare the human-machine results to those from the human-human condition (repeated here):

User	# Dialogues	# Words	Vocab	# "um"	% um
naive	48	9,315	1,076	501	5.4
expert	10	737	230	21	2.8
human-machine	86	10,622	505	380	3.6

These human-machine results appear to fall in between the naive and expert user human-human results in terms of words per dialogue, vocabulary size, and frequency of pause fillers. We suspect that this relationship between the user categories will hold for speech and grammatical false starts as well. This suggests that expert human-machine users could potentially adapt to a restricted vocabulary and still maintain efficiency. Future SLSs should plan for both the naive and the expert users.

	Total	DBQ	Reg.	First	Second
# Utterances	857	443	414	486	371
# Words	10,622	5,067	5,555	5,965	4,657
Vocabulary	505	436	505	505	435
# "um"	380	186	194	222	158
um/word (%)	3.4	3.7	3.5	3.7	3.4
% False Starts (per word total)					
Speech	0.7	0.6	0.7	0.6	0.8
Grammatical	0.9	0.9	0.9	1.0	0.8
# Error messages	219	122	97	130	89

The above table compares the database query (DBQ) with the regular condition, and the first task performed by the subject with the second task (the totals are also shown). The number of "um"s includes a variety of different pause fillers used by the subjects. The false start percentages are calculated by dividing by the total number of words observed in that session. Each subject had an average of 9 to 12 false starts per session. The number of error messages refers to the number of times subjects were presented with a "can't handle that request" response to an utterance.

In the comparison between DBQ and “regular” conditions, the only significant difference is that the “regular” condition has fewer errors than the DBQ. This suggests that the condition may not have been too constraining for the subjects; perhaps nothing that a short training session could not overcome. Differences between the first and second session, however, are larger: subjects in the first session are more verbose than in the second, and correspondingly, the first session has more error messages. These results suggest that pre-session training and user practice of the system might facilitate more efficient interaction with the machine. If one 5-minute session has this strong an effect, it is perhaps not unreasonable to consider short training sessions integrated in initial SLSs.

DISCUSSION

We found it useful to collect both human-human data and simulated human-machine data in the initial design stages of an SLS. We found that subjects could perform the air travel planning tasks when they were constrained to use only database queries, and when they were allowed a little more flexibility. Several of the subjects who started out with the DBQ condition used database queries even in the less constrained condition. Since users were familiar with database queries by the time they reached the second condition, they chose the shortest possible solution. Practice is a major factor in improving the efficiency and accuracy of completing a flight reservation, both for the human-human data and for the human-machine data.

It is important to note that subjects who believed the system was fully automated did not always use simple and clear speech. Several of the subjects said that they were impressed by the superior capability of our 'automated' system. Perhaps this overestimation of technological capability is what allowed these subjects to slip into more complex communication (larger vocabulary, more indirect requests, wandering train-of-thought utterances, more complex grammatical constructions). It is difficult to underestimate the effect of the wizard's reactions on the resulting data.

Future directions

Our data collection effort will diverge at this point. One effort will be aimed at efficient elicitation of database queries for SLS kernel evaluation. Our major effort,however, will be aimed at designing an appropriate interface for the air travel planning domain. Both efforts will involve the design and evaluation of short training sessions. We intend to run a large number of subjects on the simulation in order to assess various ideas we have about the proper interface.

User friendliness becomes more of an issue as systems become more complex and replace human-human interaction. Subjects in our human-machine experiment and subjects in other simulations (van Katwijk et al. 1979), after participating in the experiment, expressed similar frustration when the system gave a vague or inadequate error message to a multi-word and sometimes complex utterance. Subjects would like error messages to address specific reasons for rejecting an utterance: for example, inability to recognize or parse correctly, or receiving a request that the database cannot handle. It may be possible to distinguish some categories of "errors" in near-term systems, but we suggest that knowing why a request cannot be handled in many cases is nearly as difficult as handling it in the first place. Not telling the subject why a request could not be handled often results in a series of variations that have nothing to do with the real reason the request was not handled. It also causes the subjects to limit their utterances to constructions that appear to work. For these reasons, we believe it is important to consider short training sessions for subjects. Initial systems can also be constructed to mitigate the problem of the user not knowing much about the system in the same way that travel agents deal with the same problem: by taking a more active role in guiding the dialogue.

Acknowledgements

We gratefully acknowledge American Express Travel Related Services for facilitating the collection of speech data from their travel agents, the many SRI employees who agreed to have their speech recorded, the online Official Airline Guide for making their database available to us, and Steven Tepper for programming the interface and creating the tools for the "wizard". This research was funded by DARPA contract N00014-85-C-0013 and SRI International Internal Research and Development funds.

References

J. Bernstein and G. Baldwin. "Spontaneous vs prepared speech." Presented at the 110th meeting of the ASA, Nashville, TN, November, 1985.

A.F.V. van Katwijk, F.L. van Nes, H.C. Bunt, H.F. Muller & F.F. Leopold. "Naive subjects interacting with a conversing information system." *IPO Annual Progress Report,* 14:105-112, 1979.

S.L. Oviatt and P.R. Cohen. "Discourse structure and performance efficiency in interactive and noninteractive spoken modalities." Technical Note 454, Artificial Intelligence Center, SRI International, Menlo Park, California, November, 1988.